

Dear Colleagues:

A critical evaluation of:

"On the Discriminability of the Handwriting of Twins."

53(2) *Journal of Forensic Sciences*, 430-446 (March 2008)

Sargur Srihari, Chen Huang and Harish Srinivasan.

SUMMARY STATEMENT: This paper assesses the accuracy of the observations and the reliability of the protocol that the authors report for their research project into the handwriting of twins. My most critical assessment is that the authors' work has no practical application for case work. Why? Because *a computer method as they have designed can replicate a forensic document examiner's performance only if the forensic document examiner is using an inadequate methodology.*

In order for a competent expert to state to which of the several possible causes a particular handwriting feature is to be attributed, there is required both an extensive knowledge of findings from all sciences researching handwriting and a capacity to formulate and test alternative hypotheses. *The authors unwittingly prove that this requirement could never be accounted for in a computerized forensic analysis of handwriting, unless the examiner inputs it manually or takes over from the computer at the critical juncture.* Therefore, such computer testing of forensic handwriting analysis, as the authors report, is of scant to no value in establishing either the validity of theory or the reliability of method and of performance in the discipline.

On page 430, after speaking of other scientific investigations involving twins in the medical, social and biological sciences, the authors conclude: "On the other hand, handwriting is more of a behavioral characteristic, with a significant psychological component associated with it, which makes the study of the handwriting of twins to be meaningful." This is not a keynote for their research project, because behavioral or psychological factors are not reported as having been considered in formulating their method or observations. On page 446 appear two other factors, which also are not reported as having been considered in formulating their method or observations, and these are stated thus: "Distinguishing between the handwriting of twins is harder than that of nontwins because twins are more likely to have the same genetic and environmental influences than others." Nor do the authors report having substantiated either set of factors and having established their alleged links to handwriting. Unfortunately, as researchers, they do not reference nor integrate the laws, causes and dynamics of the production of handwriting into their work, in so far as we can tell from this research report.

I will go through their paper, taking statements in order and offering critique of same.

First, we must establish two principles of critical evaluation of any alleged scientific research related to our discipline that is heavily laden with unfamiliar terminology and complex equations from another discipline, in this case a computerized, algebraic statistical analysis. Our two principles are these:

1. We examine samples of handwriting, provided the authors are kind enough to share some with us. If, based on what is provided to us, their reported data are off target or even erroneous, we know that their computations and conclusions cannot be valid. So we ask ourselves: Do the samples reproduced support the mathematical evaluations and inferences based on them?
2. We then focus on the bottom line, the results offered as practical support for our work in the field. In this case their bottom line is purported proof that what forensic document examiners do in handwriting comparisons and identification is reliable. So we ask ourselves: Do the authors truly address what we do, and does their methodology reflect ours?

On both scores we will see that the authors fail us.

On page 431 they define the three things they do: "Verification is the task of determining whether a pair of handwriting samples was written by the same individual. Identification is to find the writer having the closest match with the questioned document out of a pool of writers. Recognition is to convert images to text." We must not assume that, because experts in another discipline use the same terms we do, the terms mean the same for them as they do for us.

We can easily recognize the third definition as the standard meaning of "recognition" in the computer's function of "optical character recognition." The other two definitions, for "verification" and "identification," need some clarification in case a reader might uncritically think they amount to the same thing in our parlance. For us, "identification" is to offer an opinion as to who wrote a questioned writing, while "elimination" is to offer an opinion as to who did not write a questioned writing. For us "verification" is to ascertain whether or not we are correct in a statement we have made about any component of our work. For example, we verify an observation of fact by rechecking the fact, we verify a theory by having reference to the authorities in the field, and so on for any other aspect of our work.

"Verification," as the authors define it, is to say that an individual was the one who wrote two samples, one being the individual's known sample and the other being questioned. As stated above, in our discipline that is really to make an identification. Contrariwise, "identification," as the authors define it, is to say which individual of several most likely wrote two writings, one being the questioned writing and one being the individual's

known sample. As stated above, that is really an identification of the one selected as the writer and elimination of the ones not selected as the writer. Even as described by the authors, the two allegedly different processes do not follow distinctly different methods, nor do they achieve distinctly different results. In their verification, the one suspect's exemplar is compared with the questioned writing until it can be determined whether or not this suspect can be identified as the writer of both. In identification, it is determined which of several individuals, if any, wrote the questioned writing, after comparing each individual's exemplar with the questioned writing. They use "paired with" for "compared with."

It is obligatory of any author, who wishes to research or test another discipline's reliability, to understand such discipline's theories and methods and employ its proper terminology. Thus, the authors are excused only if they were misled by document examiners who themselves had not mastered their craft and terminology.

There are two things that the authors did not seem to recognize. First, their definitions of "verification" and "identification" are the same difference, as children used to say when I was in first grade. Second, in correct forensic practice a handwriting identification is based on establishing a set of significant similarities without any unexplained significant differences. The authors never address significance for any of their observations, and they never address the need to give a reasonable explanation for any significant difference. The reasons are simple. One, they did not first state what a forensic handwriting identification is, and, two, in their academic culture they perforce must run things through complex computer equations. For example, on page 435 they give:

"The LLR in this case has the form

$$LLR = \sum_{i=1}^c \sum_j \sum_k [\ln p_s(d_i(j, k)) - \ln p_d(d_i(j, k))] \quad ."$$

Equations such as this certainly are such as no handwriting examiner has or ever would run, for good reasons that would be too lengthy to go into at the present. However, a computer cannot supply a lack of knowledge and skill in performing a comparative analysis, while the computerized complexity of data, equations, charts, graphs, and so on in the authors' paper only distracts from the several invalidities of their methodology.

Still on page 431, I quote two paragraphs to illustrate that their methodology is not the standard, generally accepted methodology in forensic handwriting examination:

"Statistical parameters for writer verification are built into CEDAR-FOX, which were obtained using several pairs of documents, which were either written by the same writer or by different writers. Writer verification consists of four [sic] steps: (1) writing element

extraction, (2) similarity computation, (3) estimating conditional probability density estimates for the difference being from the same writer or from different writers (as Gaussian or Gamma).

"Given a new pair of documents, verification is performed as follows: (1) writing element extraction, (2) similarity computation, (3) determining the log-likelihood ratio (LLR) from the estimated conditional probability density estimates."

"LLR" appears often subsequently. From its usage by the authors, I believe it means the chances of a so-called verification being correct, but for our purposes there is no need to figure it out with exactitude, since it is the end product of very poor observations, that I will address below. Thus, in the words of the old adage, it is the computational "garbage out" of the data "garbage in." In any critical evaluation of this sort, we focus on the very purpose of the project and its bottom line results in light of that purpose.

Also on page 431, the authors state regarding the computer method they designed: "The system computes three types of features—macro features at the document level, micro features at the character level, and style features from the bi-grams and words. Each of these features contributes to the final result to provide a confidence measure of whether two documents under consideration are from the same or different writers."

"Style" for them does not mean "style" for us, the latter being the master pattern one habitually employs, derived from a learned or an original model. In fact, I must confess I am not sure what "style" means for them. However, the artificial system of observations they created has no relation to the criteria for either the standard for identification of a suspected writer or the standard for the elimination of a suspected writer. I assume they are all brilliant in their proper fields, but in putting their specialty to the service of another specialty, they should employ the standard terminology of the field they are serving and design their complex computer operations to follow this specialty's methods and standards. Nothing in their article indicates any such adaptation, and, as they describe it, they are testing a non-standard method, not what is generally accepted in the discipline and what has been received in courts of law as reliable both before and after Daubert.

Page 432 is a full-page illustration of three grayscale images of letter "e," two by the same writer and one by a different writer. Four sets of impressive computer-generated imagery purport to prove the fact of same/different writer. However, the astute observer of handwriting sees immediately that the third grayscale "e" is cut off, most probably from its connection to the next letter. There is the possibility that it is also cut off from a connecting stroke from a preceding letter. It is a rudimentary principle of handwriting dynamics that an individual will most probably write the same letter observably different

depending whether it connects or does not connect to neighboring letters, whether it is in an initial, medial or terminal position in a word, or is in isolation.

We skip the next five pages that are heavy into the esoteric processes of computerized processes and statistical calculations. Since the test of the pudding is in the taste, we look for the next bite of laboratory-produced research pudding. Page 438 is the full-page Figure 9, the legend reading: "Similar handwriting samples of a pair of twins: (a) Twin 003a and (b) Twin 003b. The LLR value between these two documents is 7.15."

First, they neglected some basics in reproducing sample writings. They gave no indication of ratio of the reproduction to the original. The reader is left to assume both samples are of the same ratio to their originals. Did the writer have a limited space for the text written? If so, what effect did that have on spacing and expansion? If any effect, the validity of the samples for comparative purposes is severely reduced. What are the sex, age and handedness of these two writers? They do not tell us. They do claim that no writer had a physical condition that would affect the writing, but, given no display of an adequate knowledge of the laws and dynamics of the production of handwriting and of the criteria for its identification, such claim is questionable. The positive LLR 7.15 number indicates the likelihood of not being able to distinguish the two writers, while a negative number indicates the opposite. An examiner of modest competence can observe several significant differences between the two samples, and so conclude to a definite elimination of either twin as being the writer of both samples. If presented in a court of law as support for an expert opinion, this research report would be impeached by Figure 9 alone.

There is another important point Figure 9 demonstrates. The two writers in Figure 9 only demonstrate having learned the same school model, so that all similarities are merely class characteristics. The method the authors developed focuses on those features most related to class characteristics based on the school model.

On page 439 are more samples from two twins whose samples have a -98.36 LLR score. Here the authors' computerized method of observation accidentally serves them well. Although the two writings clearly follow the same school model, one is characterized by left slant and simplification, the other by right slant and amplification. The second writing also shows a physiological difficulty in making lower zone forms. If the second sample were in an actual case, the competent document examiner would want to ascertain whether or not the writer had a physical impairment that might affect handwriting, which is one of the possible causes of the awkwardness that should be accounted for by the handwriting expert who truly has expertise. I believe this requirement could never be

accounted for in a computerized forensic analysis of handwriting, unless the examiner inputs it manually or takes over from the computer at the critical juncture.

Two tables on page 440 show how their method validates what is not valid. They put all the data, both similarities and differences, into a computational Mulligan's stew and serve up the amalgam as a scientific product. That kind of mathematical computation is not valid in handwriting identification. The correct relationship between similarities and differences in handwriting comparison is not employed in their calculation. As Ordway Hilton said, one unexplained significant difference prevents us from making an identification, and thus it outweighs any number of significant similarities. The authors give no indication of the need to take into consideration what makes a feature significant for identification, nor is the concept "significant for identification" mentioned. Based on this report, the computer's mechanization of image recognition and its statistical computations is not designed to allow for such essential considerations.

Also on page 440 they say: "The outcome of a writer verification experiment can be one of four possibilities," namely, true positive (a correct identification), true negative (a correct elimination), false positive (an incorrect identification), and false negative (an incorrect elimination). Are they aware of a critically important possibility the competent examiner must always be alert to: Is the writing under examination fit material for a forensic examination? Not having taken a critically important possibility into consideration, they cannot reach a reliable outcome. This lack in their system may come from unawareness or from incapacity of a computer system to allow for what most makes forensic handwriting examination an objective exercise founded on scientific discoveries about handwriting. Either way, this system is an inadequate and invalid method of handwriting observation and comparison.

Still on page 440, they discuss the rate of error. If the samples on pages 438 and 439 are typical of twin samples that are close to and far from each other, there ought not have been any error rate in distinguishing twin writings. So, as readers, we must be charitable and assume some pairs of twin writings had scant significant differences, if any. Otherwise, we must assume all of the observers used were almost as unaware and inept as the computer system. The overall error rate for twins was 12.6% and for nontwins 3.15%. I submit that the nontwin matches were not comparable to the twin matches unless the two nontwin writers learned the same school model. Later they note that the error rate for identical twins was greater than for nonidentical twins. We cannot draw any inference from these error rates until we are able to factor in which twins had the cultural conditioning of being made to dress the same, act the same, in short be forced into being psychologically, socially and behaviorally joined twins, and we would need to know the

impact of such forced imposition has upon a child's psyche and behavior versus a truly voluntary experience of that sort.

Page 440 ends with an introduction of how results can be made to look better: Just fudge! They set an arbitrary rejection rate for observer reports. As the parameters for acceptance are constricted, error rates drop. This is the first research paper of any type that I have read where the authors were so candid and forthright as to explain how an error rate could be made to look poor or good or better. It leaves one with skepticism about all purported research into rates of error. Maybe only disinterested researchers, neither involved with nor known as challenging the practitioners under scrutiny, should be selected for such research. Further, if either document examiners helped determine who would do the research or only document examiners of one self-interested segment of the discipline were assisting the researchers, the entire process may have been vitiated ab initio.

On page 445 they discuss comparative error rates among a group of three questioned document students, the computer system, and a group of nine graduate students in computer science and engineering who were the lay observers. The threesome did better than the computer, which did better than the nine. The results for nontwin writings, about 95% accuracy, matched the reported accuracy from the previous study out of the same lab, which was marked by the same inadequacies in theory, method and observation. An uncritical acceptance of both studies permits one to claim that reliability has been established by replicated scientific research.

Page 446 brings us to the end of the research report. The bibliography first lists five books: Osborn's major text, Edna Robertson's book (giving at least one delightful surprise in all of this), Bradford and Bradford, Huber and Headrick, and Ordway Hilton's genuine second (revised) edition, not the alleged posthumous, second edition of 2006. Had the authors carefully studied Hilton, they would have known, and could have addressed, the inadequacies of their research method. A computer system as they have designed can replicate a forensic document examiner's performance only if the forensic document examiner is using an inadequate methodology. Unfortunately, we must beware that such examiners might now point to purported scientific proof that as forensic witnesses they are reliable. The rest of us must steadfastly hold to reality, however less impressive it is as compared to complex computational fabrications and to research projects of this type.

There are possible virtues in this research report. Designers of future research projects may be inspired to conduct a better study of the applicable literature. Future protocols can be much better designed. The information on twin writing might have value outside of the failed forensic purposes of this study. That students of document examination and not

experienced practitioners were used as the professional subjects might conceivably protect the discipline from the negative aspects of the study. If only one segment of the discipline presently has direct relations with these and similar researchers, the outsiders are protected from the implications of the poor forensic practices represented. In the future, researchers might have the better judgement, scientific objectivity and good social graces to extend a hand to all segments of the discipline.

I wish to thank Jacqueline Joseph, CDE, of Portland Oregon, for her critical evaluation of this critical evaluation. The paper was greatly improved, while, as all authors will confess, remaining faults are mine alone. For her professional services, contact Ms. Joseph at [**jjoseph@jjhandwriting.com**](mailto:jjoseph@jjhandwriting.com). Ms. Joseph also distributes my professional publications, available on CD in PDF format. Available are *QDE Index* and a collection of monographs on major issues in document examination.

Respectfully submitted for your consideration,

Marcel B. Matley